Euphytica (2009) 170:339–345
DOI 10.1007/s10681-009-9991-7

Correlations between visual biomass scores and forage yield in space planted red clover (*Trifolium pratense* L.) breeding nurseries

Heathcliffe Riday

Received: 11 March 2009 / Accepted: 18 June 2009 / Published online: 12 July 2009

Abstract Red clover (*Trifolium pratense* L.) forage yield remains a prime breeding target for improved variety development. In a world of decreasing forage legume breeding resources, rapidly and cheaply phenotyping plants for the highly quantitative trait of forage yield is vital. Many red clover selection programs are based on space planted nurseries. The objectives of this study were to determine: (1) the accuracy of visual forage yield scores in predicting actual forage yield; (2) the nature of the relationship between visual scores and actual measurements; and (3) The repeatability of visual scores between different evaluators. Twenty-seven halfsib families were transplanted at two locations in three replicates of six plant plots. Individual plant fresh weights and visual scores, by two evaluators, for forage yield were taken. On an individual plant basis visual forage yield scores showed an exponential relationship with actual fresh weights. Individual plant visual scores were very accurate with a pseudo-R^2 of 0.79 observed for the exponential model. On an entry mean basis using a linear model, visual scores could explain 90% of the variation of actual fresh weights. Agreement among evaluators scoring the same plants was very high with coefficients of determination at 0.84 for individual plants and as high as 0.96 on an entry mean basis. This study suggests that visual scores of plants in space planted red clover breeding nurseries are basically as accurate as measuring actual yields and that plants can be consistently scored the same by different evaluators.

Keywords Red clover · Biomass · Breeding · Scores · Yield

Mention of trade names or commercial products in this publication is solely for the purpose of providing specific information and does not imply recommendation or endorsement by the US Department of Agriculture.

H. Riday (✉)
US Dairy Forage Research Center, USDA-ARS, 1925 Linden Drive West, Madison, WI 53706, USA
e-mail: riday@wisc.edu

Introduction

Red clover (*Trifolium pratense* L.) is a major forage legume grown on approximately 4 million hectares worldwide and serves a myriad of functions in agricultural rotations. Primary breeding targets remain increased biomass yield production and persistence. Many red clover breeding programs remain world wide, however, resources to maintain these programs are often limiting. The ability to rapidly evaluate large numbers of individual plants in a red clover space planted breeding nursery saves resources and allows increased selection intensities due to the ability to evaluate larger number of plants. Biomass yield is a highly quantitative trait and subject to extensive genotype by environment interactions making field observations over multiple harvest and years essential for improvement.

Red clover selection is often accomplished in space planted nurseries to track individual plant performance. Space plant nurseries usually contain thousands of individual plants under observation. To accurately determine plant biomass the standard procedure is to harvest above ground individual plants whole and dry down the harvested material to determine grams dry matter per plant. Such protocols are very labor and time intensive. To save time and labor devoted to evaluating each individual plant red clover breeders often revert to a visual "score" or estimate of biomass yield. Such scores are often termed "plant vigor" scores and in addition to biomass yield try to estimate general desirability of the plant as a whole. The labor and time savings in using a score versus an actual measurement is easily a factor of 10.

The efficacy of using visual estimates of grain yield traits has been explored in small grains, soybean (*Glycine max* L.) and cotton (*Gossypium hirsutum* L.) (Bowman et al. 2004). In the cases of soybeans and cotton, selecting for increased seed cotton yield using visual selection has been successful, while in small grains it has not been (Bowman et al. 2004). Visual selection for forage yield has been successful in various forage grass species such as: small grains (Atkins et al. 1969; Ud-din et al. 1993), Pensacola bahiagrass (*Paspalum notatum* var. *saure* Parodi) (Burton 1982); rye (*Secale cereale* L.) (Bruckner et al. 1991), meadow fescue (*Festuca pratensis* Huds.) (Casler and Van Santen 2000) and perennial rygrass (*Lolium perenne* L.) (Smith et al. 2001). Coefficients of determination (r^2) between visual estimates and actual forage yield varied across the various studies with Atkins et al. (1969), reporting relatively low values for oat (*Avena sativa* L.) nurseries from two separate evaluators ($r^2 = 0.10$ and 0.24) to higher values for barley nurseries ($r^2 = 0.52$ and 0.58). Burckner et al. (1991) in rye reports visual estimates correlated with actual forage yields from $r^2 = 0.67$ to 0.81 across multiple environments. In perennial ryegrass visual estimates correlated with actual forage yields range from $r^2 = 0.17$ to 0.85. Finally in meadow fescue Casler and Van Santen (2000) report correlations from $r^2 = 0.75$ to 0.94. In cases were multiple individuals were evaluating the same sets of plants correlations between evaluator estimates were high (perennial ryegrass, $r^2 = 0.80$; oats $r^2 = 0.71$ and 0.77; and rye $r^2 = 0.77$) (Atkins et al. 1969; Smith et al. 2001).

The author is unaware of studies determining the accuracy of visual forage yield estimates in red clover, particularly under space planted conditions in companionship with a dense sod forming grass. The objective of this study was to determine the accuracy, repeatability, and relationship of visually assessed forage yield used to estimate actual fresh weight yield of red clover space plants on the individual plant level to the experiment halfsib family mean level.

Materials and methods

In this study 27 halfsib families from three related red clover populations (C584, C584-ST, and C584-UN) were used. Halfsib seed from each entry was planted in the greenhouse in February 2008. Experimental entries were transplanted in replicated incomplete block designs with three replicates per location and five incomplete blocks per replicate at: US Dairy Forage Research Dairy Farm, Prairie du Sac, Wisconsin, USA (43°21′N, 89°45′W) in a Richwood silt loam (Fine-silty, mixed, superactive, mesic Typic Argiudolls) on 23 April 2008; and University of Wisconsin-Madison, Arlington Agricultural Research Station, Wisconsin, USA (43°18′N, 89°21′W) in a Plano silt loam (fine-silty, mixed, superactive, mesic typic argiudolls) on 24 April 2008. Six plants per plot were transplanted directly into naturalized Kentucky bluegrass (*Poa pratensis* L.) pasture at Prairie du Sac and a fine fescue turf blend (*Festuca rubra*–65%, *brevipila*–20%, and *ovine*–15%) at Arlington, with plants within plots spaced 30 cm apart in a row. Six plant plots were planted in rows with 60 cm between plots, within rows, and 75 cm between rows. Space plant nurseries were mowed on 18 Jun and 15 Jul 2008 at Prairie du Sac and Jun 24 and 8 Aug 2008 at Arlington.

Nurseries were evaluated for biomass yield on 25 Aug 2008 at Prairie du Sac and 3 Sep 2008 at Arlington. During each evaluation visual scores to estimate total above ground forage yield per plant were accomplished by two people. Visual scores were given on a 0–5 scale with ¼ unit increments allowable. Although the main target was to visually assess yield, general scoring guidelines were given to evaluators. Generally plants scored one were small and had a few leaves with no stems present; plants scored three had some fully formed stems; and plants scored five were large and had many robust stems and leaves. On the same day that the nursery was visually scored individual plants were harvested at a 5 cm height and weighed fresh. A third forage yield estimate was calculated as the average of the two visual scores taken for each plant.

Results

Individual plant basis

Forage yield visual scores and fresh weights were taken under typical space plant nursery conditions. The mean plant fresh weight at Prairie du Sac was 48 g, while the mean fresh weight at Arlington was 20 g (Table 1). The visual yield scores had a mean of 2.7 for both evaluators at Prairie du Sac and 2.0 and 2.3 at Arlington for each evaluator respectively (Table 1). The Arlington nursery was scored in early September, which was later in the growing season than the Prairie du Sac measurements. The Arlington plants showed more dormancy response than the Prairie du Sac location. Due to increased dormancy response the Arlington plants were smaller and appeared visually to have less variation and were more difficult to visually score.

Yield differences between harvests and locations can be even greater than observed in this study. To account for unequal variances between nursery locations and harvests, as well as variance differences between weight and score units, the four data sets (2 visual scores, average visual score, and actual fresh weights) were normalized for each replicate within each location. Normalized Fresh weight ranges and distributions were positively skewed at both locations (Table 1). Normalized visual scores tended to be more normally distributed than the fresh weights and have a tendency to be negatively skewed. In previous unpublished observations the author has noted that space plant nurseries placed into dense grass sod pastures tend to have visibly notable micro-field variation (Riday, unpublished). To remove this field variation plots in this study were divided into incomplete blocks within replicates. The fixed incomplete block effects were removed from the normalized data sets. Incomplete block effects were greater at Arlington as compared to Prairie du Sac (Table 1). Blocking had a similar effect on fresh weights and visual scores and in this study accounted for approximately 10% of the total normalized data variances.

One of the core questions of this study was to determine the relationship between visual scores and fresh weights. Having removed all fixed environmental effects (i.e. location, replicate, and block effects) and transformed both yield measures to the same scale a plot of the normalized score average yield versus normalized plant fresh weight reveals a exponential relationship (Fig. 1) with a pseudo-R^2 (defined as $1 - SS_{residual}/SS_{corrected\ total}$) of 0.79. This suggest that the human mind can accurately estimate single plant red clover yield but it does so by orders of magnitude rather than in a linear fashion (i.e. this plant is X times as big as that plant and that plant is X times as big as the next plant etc.). This observation dovetails nicely with the differing distribution skewness of the visual scores and actual fresh weight measures. Whereas actual yield distributions tend to be compressed towards zero with a few exceptional plants forming the skewed right tail of the distribution; the visual scoring metric essentially is on a logarithmic scale.

The residual plot of the exponential relationship between visual yield scores and actual fresh weight reveals that the larger the relative size of the plant the more difficult it becomes to accurately measure it (Fig. 2). Evidence for this conclusion comes from the positive regression slope (slope = 0.09; $P < 0.0001$) of normalized forage yield score against the absolute value of the residual being positive (Fig. 2).

Both evaluators were equally successful at visually predicting yield, although better predictions were made at Arlington compared with Prairie du Sac (Table 2). Averaging the score of both evaluators increased the prediction efficiency slightly (Table 2). Correlations on an individual plant basis between the two visual scores was very high at $R^2 = 0.88$ and 0.84 at Arlington and Prairie du Sac respectively. These results suggest that in a resource limited environment a single visual score of each plant during each rating period is sufficient to achieve very good plant yield estimates.

Table 1 Fresh Weight, score 1, score 2, and score average: means, standard deviations, minimums, maximums, and skewness for unadjusted plant data, normalized plant data, normalized plant data adjusted for incomplete blocks, normalized plot mean data, normalized entry mean data on a location basis, and normalized entry mean data from evaluations made at two Wisconsin, USA locations

Biomass yield measure	Unadjusted plant data							
	Arlington				Prairie du Sac			
	Mean	SD	Min.	Max.	Mean	SD	Min.	Max.
Fresh weight[a]	19.96	14.34	0.00	84.90	48.11	34.24	0.70	227.80
Score 1[b]	1.98	0.83	0.00	4.25	2.72	0.79	0.25	5.50
Score 2[b]	2.27	0.74	0.00	4.00	2.73	0.82	0.10	5.00
Score average[b]	2.13	0.76	0.00	4.00	2.72	0.79	0.18	5.00
	SD	Min.	Max.	Skew	SD	Min.	Max.	Skew
	Normalized plant data[c]							
Fresh weight	1.00	−1.43	4.54	0.93	1.00	−1.54	5.61	1.20
Score 1	1.00	−2.58	2.71	−0.14	1.00	−3.13	3.34	−0.52
Score 2	1.00	−3.06	2.48	−0.81	1.00	−2.97	2.85	−0.33
Score average	1.00	−2.83	2.64	−0.50	1.00	−3.01	2.93	−0.48
	Normalized plant data adjusted for incomplete blocks							
Fresh weight	0.92	−1.92	4.10	0.77	0.98	−1.68	5.32	1.15
Score 1	0.90	−2.98	2.51	−0.32	0.97	−2.99	3.08	−0.45
Score 2	0.95	−3.29	2.27	−0.75	0.98	−2.87	2.68	−0.21
Score average	0.92	−3.22	2.43	−0.57	0.98	−2.86	2.76	−0.37
	Normalized plot mean data							
Fresh weight	0.48	−1.29	2.21	0.78	0.59	−1.21	1.53	0.27
Score 1	0.46	−1.93	1.58	−0.39	0.56	−1.35	1.34	−0.10
Score 2	0.46	−2.07	1.30	−0.92	0.54	−1.37	1.11	−0.26
Score average	0.46	−2.05	1.47	−0.69	0.56	−1.37	1.20	−0.20
	Normalized entry mean data per location							
Fresh weight	0.34	−0.71	0.82	0.35	0.44	−0.92	1.00	0.37
Score 1	0.31	−0.57	0.64	−0.41	0.41	−0.86	0.70	−0.62
Score 2	0.30	−0.59	0.60	−0.53	0.41	−0.82	0.70	−0.58
Score average	0.30	−0.56	0.63	−0.45	0.41	−0.85	0.67	−0.64
	Normalized entry mean data							
Fresh weight	0.35	−0.74	0.69	−0.35				
Score 1	0.32	−0.69	0.67	−0.17				
Score 2	0.31	−0.71	0.57	−0.34				
Score average	0.32	−0.71	0.62	−0.27				

[a] Grams plant^{-1}

[b] Visual forage yield score: 0–none to 5–lots

[c] Normalized data transformation $= (x - \bar{x})/\text{std. dev.}$

Plot mean basis

Plot means are important in halfsib family selection. For biomass yield halfsib family plots can be bulk harvested, dried, and weighed, with the dry weight divided by the number of plants in the plot. This procedure is much less labor and resource intensive than weighing each plant but is still much more

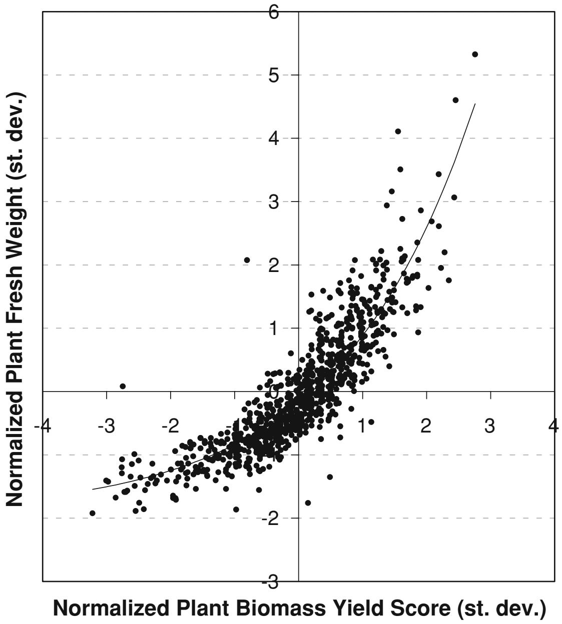

Fig. 1 Individual normalized plant fresh weights (standard deviations) plotted against the average of two normalized biomass yield scores (standard deviations) with fitted exponential function (fresh weight $= 0.3784e^{0.475 \times \text{score}} + 1.923$) (pseudo-$R^2 = 0.79$)

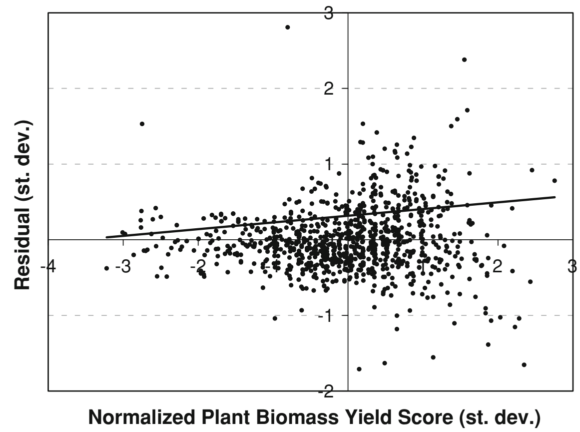

Fig. 2 Residual (standard deviations) of individual normalized plant fresh weights with the average of two normalized biomass yield scores plotted against the average of two normalized biomass yield scores (standard deviations) with a linear regression fitted between the absolute value of the residual (standard deviations) and the average of tow normalized biomass yield scores (standard deviations). The regression slope equals 0.09 ± 0.01 and is significant at $P < 0.0001$

intensive than averaging individual plant visual scores within each plot. A lot of variation is lost when working on a plot mean basis (within plot environmental error as well as ¾ of the heritable additive genetic variance). In this study approximately 65% of the variance is found within plots. The normalized plot mean fresh weights had a big reduction in yield range at both locations and still were slightly skewed at the Arlington location (Table 1). The visual scores showed similar reduction in range and reductions in skewness at Prairie du Sac.

At Prairie du Sac a simple linear relationship between visual scores and actual fresh weight was the best predictor with r^2 just above 0.7 (Table 2). Efficacy of visual scores predictions on an individual plant basis or a plot mean basis were not different at Prairie du Sac. At Arlington an exponential model still gave a better fit between visual scores and actual fresh weights but the advantage was marginal and a linear model could have easily been used as well (Table 2). As on an individual plant basis visual score predictions of yield were more accurate than at Prairie du Sac. Scorer 1 was more accurate than scorer 2 at Arlington (Table 2). Averaging the two scores provided no advantage on a plot mean basis with agreement between scores very high at both locations (Table 2).

Table 2 Coefficients of determination for normalized incomplete block adjusted visual scores of biomass yield as predictors of normalized incomplete block adjusted fresh weight and coefficients of determination between two normalized incomplete block adjusted visual scores for biomass yield taken, at two Wisconsin, USA locations

	Mean basis			
	Plant	Plot	Location	Entry
Coefficient of determination (R^2)				
Fresh weight with				
Arlington score 1	0.79[a]	0.84[a]	0.89	0.90[b]
Arlington score 2	0.78[a]	0.77[a]	0.81	0.90[b]
Arlington score average	0.82[a]	0.84[a]	0.88	0.91[b]
Prairie du Sac score 1	0.70[a]	0.71	0.78	
Prairie du Sac score 2	0.71[a]	0.72	0.80	
Prairie du Sac score average	0.76[a]	0.73	0.80	
Between score 1 and score 2				
Arlington	0.88	0.85	0.87	0.96[b]
Prairie du Sac	0.84	0.92	0.96	

[a] Pseudo-$R^2 = (1 - SS_{err}/SS_{corrected\ total})$ based on fitted exponential function

[b] Values fore entry means across locations

Entries on a location mean basis

Examining means at the entry by location level reveals almost normal data with a somewhat reduced range from the plot mean data (Table 1). Thirteen percent of the variation is accounted for by the entry by replication interaction. Standard deviations and forage yield measure (both visual scores and fresh weights) are almost identical (Table 1). Linear models best explain the relationship between visual scores and actual yield (Table 2). Visual scores are still more accurate at the Arlington location with no advantage provided by averaging both scorers. Scorer 1 at the Arlington location appears to be slightly more accurate at predicting actual fresh weight. Agreement between the two scores at this level is close to perfect (Table 2).

Entry mean basis

Halfsib family selection is conducted on an entry mean basis. At this level visual scores are 90% accurate at predicting actual fresh weights (Table 2). This study was conducted using three replicates of each entry at each location and it is probable that increased entry replication would only increase the visual score prediction accuracy. Only three percent of the total variation is attributed to entry by location interactions. The visual score data and actual fresh weight data appear to have almost identical distributions and skewness (Table 1). Linear models offer the best fits between visual scores and actual fresh weights (Table 2). Agreement between both visual scores is almost perfect at $r^2 = 0.96$ (Table 2).

Discussion

This study demonstrates the accuracy of red clover visual forage yield scores for predicting actual fresh weights in space planted breeding nurseries. The accuracy of these visual scores is in the same range as forage quality predictions using near infrared spectroscopy. It is important to note that this study measured the accuracy of visual scores for predicting actual plant fresh weights and not actual plant fresh weight selection gains. Many previous studies examining forage yield selection gains based on visual scores often confound questions of the sward yield gains from selection among space plant plots, strip drilled plots, and sward plots.

In a red clover space plant nursery taking actual individual plant yield measurements may likely be precision "overkill." With 65% of the total variation contained within plots a significant portion of this is likely non-genetic making the prospect of precisely determining the genetic component of forage yield on an individual plant basis without clonal replication almost impossible. Although repeated measures over multiple harvests should improve precision. Yet even in the later case visual scores at each particular harvest should be able to provide the level of precision necessary to evaluated forage yield.

Although obvious upon measurement the exponential relationship between visual scores and actual fresh weight provides insight into interpreting the score and training evaluators to administer the scores. From an individual plant selection perspective plants with high average visual scores across multiple harvests would have even higher selection value in terms of actual forage yield due to the exponential scale of the visual score when treated as a normally distributed variable. The exponential nature of the visual score also explains the slightly negative skew of the visual score distribution due compression of scores at the higher yielding tail of the distribution. From a practical perspective this influences scoring strategy. Since the value of high scores is exacerbated by their exponential nature, instructing evaluators to expect higher scores to be rare may improve scoring accuracy and influence them to be more "stingy" or careful with higher scores. On the scale used in our breeding nurseries and in this study scores of 4–5 are indeed more uncommon and these plants are usually tracked more closely over harvests. Being aware that higher yielding plants are more difficult to accurately visual score should also lead to more care being taken when attempting to score these plants in order to improve scoring accuracy.

This study was conducted during late summer on spring transplanted establishment year space plant nurseries grown in competition with a grass. Maximum yield for these plants during the full life of the stand would likely not occur until the spring of the first post-establishment year. A possible concern with using visual yield scores may be that in a high yielding environment (especially without grass competition) visual scores may become less accurate.

Support for this concern is evidenced by the higher correlations between visual scores and actual forage yields at the lower yielding Arlington location compared to the higher yielding Prairie du Sac location. Another concern with using visual forage yield scores may be a lack clearly visible variation among plants in a high yielding environment, making the plants more difficult to score. However, based on the authors experience in using visual scores across multiple environments, lack of visible variation has never been a concern. The only area of difficulty encountered in visual scoring in other space plant nurseries has been cases where different populations with varying growth habits (i.e. erect vs. decumbent growth habit) were in the same nursery making scoring more difficult. This study alone, however, should give red clover breeders confidence that their visual estimates of forage yield in space plant nurseries are accurately capturing enough variation in actual forage yield for plant breeding purposes.

References

Atkins IM, Merkle OG, Pawlisch PE (1969) Visual estimates and clipping plot size for evaluating the forage potential of small grain varieties. Agron J 61:88–92

Bowman DT, Bourland FM, Meyers GO, Wallace TP, Caldwell D (2004) Visual selection for yield in cotton breeding programs. J Cotton Sci 8:26–68

Bruckner PL, Raymer PL, Burton GW (1991) Recurrent phenotypic selection for forage yield in rye. Euphytica 54: 11–17

Burton GW (1982) Improving recurrent restricted phenotypic selection increases bahiagrass forage yields. Crop Sci 22:1058–1061

Casler MD, Van Santen E (2000) Patterns of variation in a collection of meadow fescue accessions. Crop Sci 40:248–255

Smith KF, Tasneem M, Kearney GA, Reed KFM, Leonforte A (2001) Evaluation of herbage yield in a forage grass breeding program: comparison of visual rating *versus* measurement in single-row plots or swards. Aust J Exp Agric 41:1161–1166

Ud-Din N, Carver BF, Krenzer EG Jr (1993) Visual selection for forage yield in winter wheat. Crop Sci 33:41–45